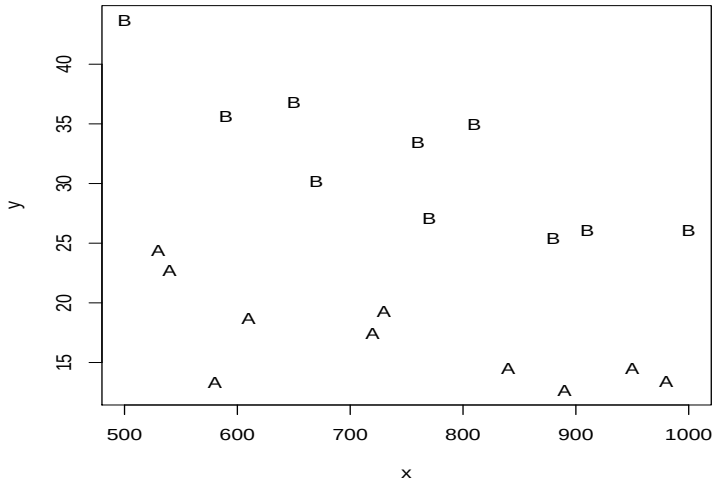
Indicator Variables

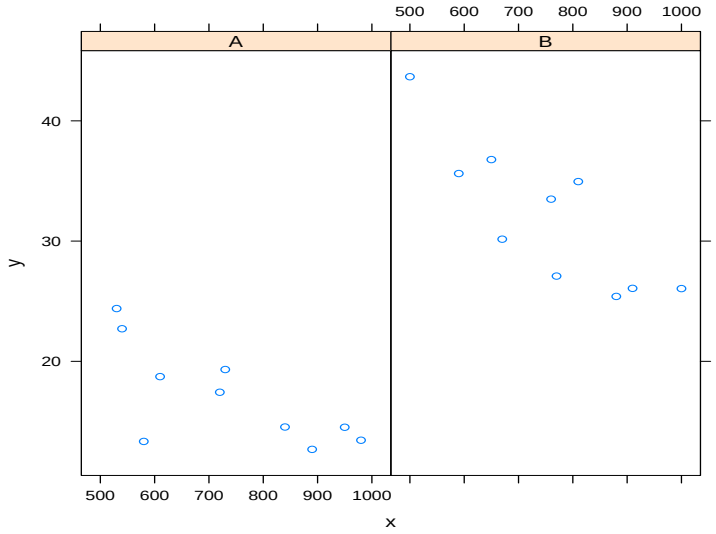
```
> tool=read.table(file="tool.txt", header=T)
> tool
```

```
    i     y    x type
1   1 18.73  610    A
2   2 14.52  950    A
3   3 17.43  720    A
4   4 14.54  840    A
5   5 13.44  980    A
6   6 24.39  530    A
7   7 13.34  580    A
8   8 22.71  540    A
9   9 12.68  890    A
10 10 19.32  730    A
11 11 30.16  670    B
12 12 27.09  770    B
13 13 25.40  880    B
14 14 26.05 1000    B
15 15 33.49  760    B
16 16 35.62  590    B
17 17 26.07  910    B
18 18 36.78  650    B
19 19 34.95  810    B
20 20 43.67  500    B
```

```
> y=tool[,2]
> x=tool[,3]
> type=tool[,4]
> plot(x,y, pch=as.character(type))
```



```
> library(lattice)
> print(xyplot(y~x|type))
```




```
> f1=lm(y~x+type)
> print(f1)

Call:
lm(formula = y ~ x + type)

Coefficients:
(Intercept)            x        typeB
   35.20873     -0.02456     15.23547

> summary(f1)

Call:
lm(formula = y ~ x + type)

Residuals:
     Min       1Q   Median       3Q      Max
-7.62550 -1.63085  0.06125  2.22184  5.50445

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 35.208726   3.738882   9.417 3.71e-08
x           -0.024557   0.004865  -5.048 9.92e-05
typeB       15.235474   1.501220  10.149 1.25e-08

Residual standard error: 3.352 on 17 degrees of freedom
Multiple R-squared: 0.8787,        Adjusted R-squared: 0.8645
F-statistic:  61.6 on 2 and 17 DF,  p-value: 1.627e-08

> anova(f1)
```

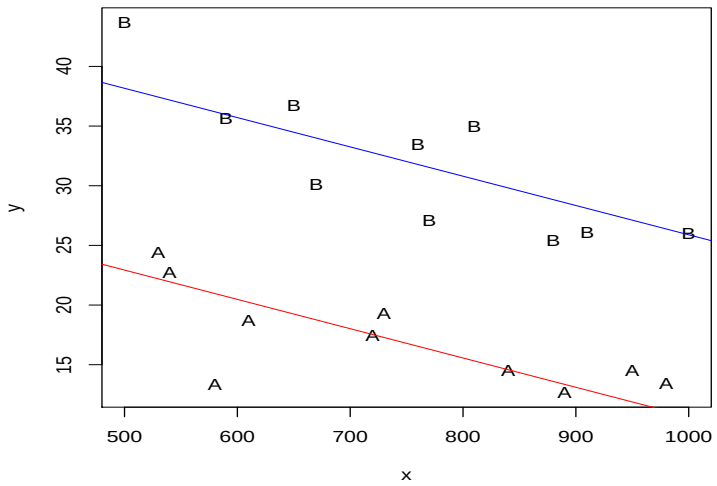
```
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)
x          1  227.03  227.03  20.209 0.0003188
type       1 1157.08 1157.08 102.997 1.246e-08
Residuals 17  190.98   11.23
```

```
> (b=coef(f1))

(Intercept)           x       typeB
35.20872630 -0.02455729 15.23547401
```

```
> (X=model.matrix(f1))

   (Intercept)    x typeB
1            1  610     0
2            1  950     0
3            1  720     0
4            1  840     0
5            1  980     0
6            1  530     0
7            1  580     0
8            1  540     0
9            1  890     0
10           1  730     0
11           1  670     1
12           1  770     1
13           1  880     1
14           1 1000     1
15           1  760     1
16           1  590     1
17           1  910     1
18           1  650     1
19           1  810     1
20           1  500     1
attr(,"assign")
[1] 0 1 2
attr(,"contrasts")
attr(,"contrasts")$type
[1] "contr.treatment"
```

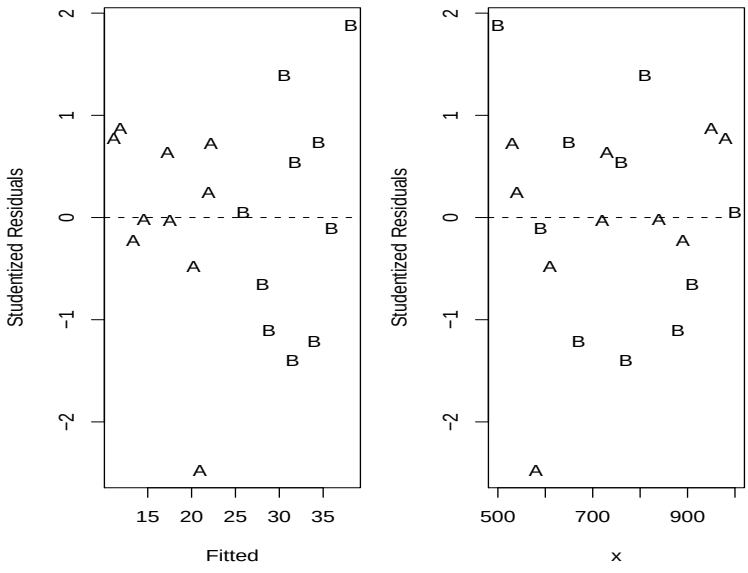
regression line:

```
> plot(x,y,pch=as.character(type))
> abline(b[1], b[2], col=2)
> abline(b[1]+b[3], b[2], col=4)
```



We are 96% confidence that for the changing from tool type A to tool type B increase the mean tool life by 12.135 to 17.873 hours

```
> par(mfrow=c(1,2),mar=c(4,4,1,1))
> plot(fitted(f1), rstandard(f1), pch=as.character(type), xlab="Fitted", ylab="Student
> abline(h=0, lty=2)
> plot(x, rstandard(f1), pch=as.character(type), xlab="x", ylab="Studentized Residuals
> abline(h=0, lty=2)
```



compare model f1 with two different regression lines

```
> x.a=x[type=="A"]
> x.b=x[type=="B"]
> y.a=y[type=="A"]
> y.b=y[type=="B"]
> g1=lm(y.a~x.a)
> anova(g1)
```

```
Analysis of Variance Table

Response: y.a
          Df Sum Sq Mean Sq F value  Pr(>F)
x.a        1 82.100  82.100  9.3177 0.01576
Residuals  8 70.489   8.811
```

```
> g2=lm(y.b~x.b)
> anova(g2)

Analysis of Variance Table

Response: y.b
          Df  Sum Sq Mean Sq F value   Pr(>F)
x.b        1 231.230 231.230  19.805 0.002139
Residuals  8  93.404  11.675
```

```
> sg1=summary(g1)
> sg2=summary(g2)
```

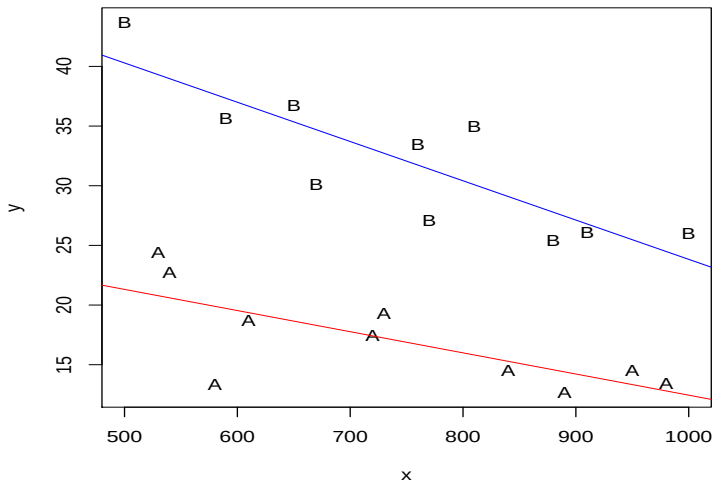
Two lines

Consider a model with different intercept and slope.

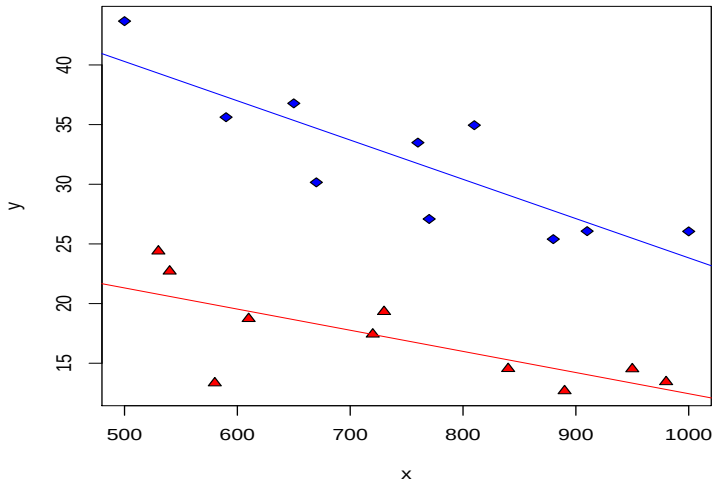
```
> f2=lm(y~x+type+x:type)  #it is the same as y~x*type
> print(f2)

Call:
lm(formula = y ~ x + type + x:type)

Coefficients:
(Intercept)            x        typeB      x:typeB
   30.17601     -0.01773     26.56934     -0.01519
```

```
> summary(f2)

Call:
lm(formula = y ~ x + type + x:type)

Residuals:
    Min      1Q  Median      3Q     Max
-6.5534 -1.7088  0.3283  2.0913  4.8652

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.176013   4.724895   6.387 9.01e-06
x           -0.017729   0.006262  -2.831  0.01204
typeB       26.569340   7.115681   3.734  0.00181
x:typeB     -0.015186   0.009338  -1.626  0.12345

Residual standard error: 3.201 on 16 degrees of freedom
Multiple R-squared: 0.8959,        Adjusted R-squared: 0.8764
F-statistic: 45.92 on 3 and 16 DF,  p-value: 4.37e-08
```

```
> anova(f2)

Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq  F value    Pr(>F)
x          1  227.03  227.03  22.1640  0.000237
type       1 1157.08 1157.08 112.9591 1.169e-08
x:type     1   27.09   27.09   2.6443  0.123451
Residuals 16  163.89   10.24

> (b=coef(f2))

(Intercept)           x       typeB     x:typeB
30.17601317 -0.01772865 26.56934009 -0.01518561

> (X=model.matrix(f2))

   (Intercept)    x typeB x:typeB
1            1  610     0       0
2            1  950     0       0
3            1  720     0       0
4            1  840     0       0
5            1  980     0       0
6            1  530     0       0
7            1  580     0       0
8            1  540     0       0
9            1  890     0       0
10           1  730     0       0
11           1  670     1     670
12           1  770     1     770
13           1  880     1     880
14           1 1000     1    1000
15           1  760     1     760
16           1  590     1     590
17           1  910     1     910
18           1  650     1     650
19           1  810     1     810
20           1  500     1     500
attr(,"assign")
[1] 0 1 2 3
attr(,"contrasts")
attr(,"contrasts")$type
[1] "contr.treatment"

> plot(x,y,pch=as.character(type))
> abline(b[1], b[2], col=2)
> abline(b[1]+b[3], b[2]+b[4], col=4)
```




```
> plot(x,y, type="n", axes=F);
> box()
> axis(2)
> axis(1, xlim=seq(500,1000, 1000))
> points(x.a, y.a,pch=24, bg=2)
> points(x.b, y.b, pch=23, bg=4)
> abline(b[1], b[2], col=2)
> abline(b[1]+b[3], b[2]+b[4], col=4)
```



use model x*type=x+type+x:type

```
> f2=lm(y~x*type)
> summary(f2)

Call:
lm(formula = y ~ x * type)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-6.5534 -1.7088  0.3283  2.0913  4.8652

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 30.176013   4.724895   6.387 9.01e-06
x           -0.017729   0.006262  -2.831  0.01204
typeB       26.569340   7.115681   3.734  0.00181
x:typeB     -0.015186   0.009338  -1.626  0.12345

Residual standard error: 3.201 on 16 degrees of freedom
Multiple R-squared: 0.8959,        Adjusted R-squared: 0.8764
F-statistic: 45.92 on 3 and 16 DF,  p-value: 4.37e-08
```

Testing for same intercept and same slope

```
> f0=lm(y~x)
> summary(f0)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-14.7931  -6.6339  -0.8645   7.2428  13.7900

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 40.79865    9.54829   4.273 0.000458
x           -0.02184    0.01254  -1.741 0.098729

Residual standard error: 8.654 on 18 degrees of freedom
Multiple R-squared: 0.1441,        Adjusted R-squared: 0.09659
F-statistic: 3.031 on 1 and 18 DF,  p-value: 0.09873

> anova(f0,f2)

Analysis of Variance Table

Model 1: y ~ x
Model 2: y ~ x * type
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     18 1348.06
2     16  163.89  2   1184.16 57.802 4.773e-08
```

Testing for same slope

```
> anova(f1,f2)
```

```
Analysis of Variance Table

Model 1: y ~ x + type
Model 2: y ~ x * type
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     17 190.980
2     16 163.893  1    27.087 2.6443 0.1235
```

Change refencec level

```
> tool$type=relevel(tool$type, ref="B")
> summary(lm(y~x+type, data=tool))

Call:
lm(formula = y ~ x + type, data = tool)

Residuals:
     Min       1Q   Median       3Q      Max
-7.62550 -1.63085  0.06125  2.22184  5.50445

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  50.444200   3.818266  13.211 2.28e-10
x            -0.024557   0.004865  -5.048 9.92e-05
typeA       -15.235474   1.501220 -10.149 1.25e-08

Residual standard error: 3.352 on 17 degrees of freedom
Multiple R-squared: 0.8787,        Adjusted R-squared: 0.8645
F-statistic:  61.6 on 2 and 17 DF,  p-value: 1.627e-08
```

More than two levels

data: Prestige of Canadian Occupations in package "car".

Description The Prestige data frame has 102 rows and 6 columns. The observations are occupations.

Usage Prestige

Format This data frame contains the following columns:

education Average education of occupational incumbents, years, in 1971.

income Average income of incumbents, dollars, in 1971.

women Percentage of incumbents who are women.

prestige Pineo-Porter prestige score for occupation, from a social survey conducted in the mid-1960s.

census Canadian Census occupational code.

type Type of occupation. A factor with levels (note: out of order): bc, Blue Collar; prof, Professional, Managerial, and Technical; wc, White Collar. Source

```
> library(car)
> print(xyplot(prestige~type, data=Prestige))
> fm1=lm(prestige~type, data=Prestige)
> anova(fm1)
```

```
Analysis of Variance Table

Response: prestige
          Df  Sum Sq Mean Sq F value    Pr(>F)
type       2 19775.6  9887.8  109.59 < 2.2e-16
Residuals 95  8571.3    90.2

> aov1=aov(prestige~type, data=Prestige)
> coef(aov1)

(Intercept)    typeprof      typewc
  35.527273   32.321114    6.716206

> summary(fm1)

Call:
lm(formula = prestige ~ type, data = Prestige)

Residuals:
      Min        1Q    Median        3Q       Max
-18.22727  -7.17727  -0.08538   6.11745  25.25652

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   35.527      1.432  24.810  < 2e-16
typeprof      32.321      2.227  14.511  < 2e-16
typewc         6.716      2.444   2.748  0.00718

Residual standard error: 9.499 on 95 degrees of freedom
  (4 observations deleted due to missingness)
Multiple R-squared: 0.6976,        Adjusted R-squared: 0.6913
F-statistic: 109.6 on 2 and 95 DF,  p-value: < 2.2e-16
```